Response to Yellowstone River Peer Review Questions

General Comments

This is a well written report on "Using a computer model to derive numeric nutrient criteria." There are relatively few errors in the draft, which made reviewing clear. The use of multiple sources of information, including a computer model, is a very good idea for establishing nutrient criteria. The many concepts developed and employed in this effort are innovative, well founded, and sound. However, I disagree with the conclusions that model conditions warrant more credibility than other sources of information and that model results should be used to set nutrient criteria for the Yellowstone River.

In summary, my short responses to the questions are:
1. The data used to run, calibrate, and validate the model were appropriate, but not sufficient.
2. Model calibration and validation were not good, because the fit of data to model runs was poor for a key endpoint variable, benthic algal biomass, and many results were biased.
3. The uncertainty of model predictions was problematic because: the model was not validated well for a key endpoint variable; the model was used to extrapolate to nutrient conditions outside the range for which it was calibrated and validated; and the model did not simulate extreme values well.
4. pH and algal biomass response endpoints should be used to establish nutrient criteria. The most sensitive response to a stressor (i.e. nutrients in this case) should be used to establish stressor criteria, even if different response endpoints are most sensitive in different types of habitats (in this case shallow and deep river habitats).
5. The appropriate methods were used to gather information about the development of nutrient criteria, but the results of the computer model were overstated and overweighted in a premature decision on nutrient criteria.

1. Please evaluate the sufficiency and appropriateness of the data used to run the model.

The data used to run, calibrate, and validate the model were appropriate, but not sufficient.

The computer model was designed to measure important response variables, such as benthic algal biomass, pH, and DO. These parameters respond either directly or indirectly to variation in nutrient concentrations and are used in either narrative or numeric water quality criteria in many states. These variables are highly appropriate from the perspective that we want to protect uses of waters. We know enough about nutrients to know the effects of nutrients instream and downstream. With proper research and synthesis of results, we should be able to set nutrient criteria above minimally disturbed conditions without threatening designated uses, such as drinking water, recreational uses and aesthetics, and support of biodiversity. Although we may not be protecting aquatic biodiversity of taxa that are highly sensitive to moderately increased nutrient concentrations in a habitat with nutrients above minimally disturbed condition, presumably those taxa are being protected in other habitats in which minimally disturbed condition is being protected (invoking tiered aquatic life uses). With the knowledge that biodiversity of some nutrient sensitive taxa will not be protected at nutrient concentrations that

generate algal biomasses greater than 150 mg chl a m^{-2} and pH and DO standard violations, benthic algal biomass, DO, and pH can be appropriate endpoints for managing nutrients.

The right variables were modeled, measured, and calibrated in the field, but the sample size was low. Many of the key environmental variables were measured in the field, but they were measured at less than 10 locations. This limits the power of the comparison, much as a low sample size limits the statistical power in hypothesis testing. Was the fit or the lack of fit of the model to data due to chance or was it true?

The study should have been designed to have the calibration and validation datasets at the same time of year, perhaps sampling during summers of 2007 and 2008. The differences in temperature and light (day length and sun angle) between August and September could be substantial given they are within range that macroalgae like *Cladophora* are especially sensitive. August and September also have very different algal accumulation histories and processes regulating algal ecology probably differ as a result. Interannual variation in physical and chemical conditions in the Yellowstone River are relatively predictable, because of discharge regulation by snowpack melting, compared to rivers in parts of the country where unpredictable rain events have great effects on discharge and resulting physical and chemical conditions (e.g. light and nutrient concentrations).

Another concern was having sufficient scientific foundation for model coefficients. Admittedly, some knowledge is better than none, but assuming that coefficients developed in lakes or other parts of the country and for different kinds of algae in one condition or another would apply to this location seems premature. Many of the parameters were developed in the 1970s or earlier, not that old is necessarily bad, but it is an indication that few new components were available or were found in the literature for use in the computer model. More field and laboratory research is needed to quantify the parameters being used in processed based models.

2. Please evaluate model calibration and validation.
Model calibration and validation were not good, because the fit of data to model runs was poor for a key endpoint variable, benthic algal biomass, and many results were biased.

Not much change was needed in many model parameters to calibrate the model, but many parameters for benthic algal growth were substantially different between the initial estimate and calibrated value (Tables 9-5, 9-6, and 9-7). Almost no discussion followed on the magnitude of these changes and if they were reasonable.

At least one set of the changes in parameters was relatively easy to evaluate and determine if they were reasonable. The mass ratio of N:P in algal cells is assumed to be 7:1, and in the Yellowstone River was often lower because of the relatively low supply of N versus P in the river. The initial mg N and P per mg algae (subsistence quotas for N and P) for benthic algae were assumed to be 0.7 and 0.1, respectively (Table 9-6).

- The real issue is the relatively large change in one value during calibration and the unrealistic ratio for parameter values resulting from that calibration. The resulting calibration values of parameters for subsistence quotas for N and P were 3.20 mg N and 0.13 mg P, respectively. Even though each of these parameters independently fit within

the range of possible values reported in the literature (remembering that one outlier in the literature has great effects on this range), the ratio seems very high for conditions within the Yellowstone River. The resulting mass ratio of subsistence levels of N and P was 3.20:0.13, which is more than 3 times the expected 7:1 ratio and 6 times the 4:1 ratios observed in low N habitats like the Yellowstone.

- Although internal N and P half-saturation constants are substantially different types of parameters than subsistence quotas, both are involved with algal growth, both were changed substantially during calibration, and ratios for both were unusually high.
- The same kinds of problems were noted for the phytoplankton (Table 9-7).
- A confusing issue initial parameter values (e.g. 0.7 mg N or 0.1 mg P per mg algae) indicate 70 and 10% of the algae were composed of N and P. Most of algal mass is carbon, not N or P. Presumably the units or my understanding of what these parameters mean were wrong.

Fit of the model, similarity between predicted and observed conditions, was better for physical than chemical parameters, and better for chemical than biological parameters. QAPP criteria were not met for 1 out of 5 of the parameters assessed (Table 10-1). The variable with poor fit based on RMSE and RE was benthic algal biomass, either by using the Q2K or AT2K model. Since benthic algal biomass was a key response endpoint, and an endpoint for which nutrient criteria were eventually going to be made, it was important that the model predict benthic algal biomass well.

As suggested on page 10-21, I agree that the AT2K model "allows us the ability to gain better information about spatial relationship of biomasses across a river transect," but I don't agree that AT2K model predictions were sufficiently accurate for the purposes intended for the modeling effort. High benthic algal biomasses were consistently under-predicted.

During review of figures, I became concerned that deviations between observed conditions and conditions predicted by the model are more serious if they are biased than if they are randomly distributed above and below model predictions. This bias would not be captured in the RMSE and RE statistics for goodness of fit. For example, even though the RE is only 7.3% for TN calibration and 1.38% for validation (Figure 10-7, the model overestimates TN concentrations). The bias in predictions (residual error) is common in many of the nutrient and biological parameters. In most cases, bias was either high or low along the river, but in some cases it systematically switched from high to low, which you could imagine was the case for the August 2000 phytoplankton validation (Figure 11-9). Systematic bias along the river is a concern because habitat conditions change systematically along the river.

The model did not capture extreme conditions well, especially for benthic algae. If there was little variation, the model tended to fit much better than if a parameter varied greatly over the range of nutrient and habitat conditions in the river. For example, diurnal variation in dissolved oxygen and discharge were simulated well by the model, but pH and benthic algal biomass which varied much more than DO and discharge were not simulated well by the model.

The model may not have been able to simulate the high algal biomasses that accumulate in the river. For example in Figure 10-15, the model never predicted algal biomass to be greater than

about 70 mg chl a m^{-2}. However, several observations of higher chlorophyll were observed. In addition, most of the observed levels of chlorophyll a were less than 50 mg chl a m^{-2} and fell within a confidence envelop that probably had a width of 40 mg chl a m^{-2}. So it would have been difficult for the model to be wrong when benthic algal biomass was less than 50 mg chl a m^{-2}. When benthic algal biomass was predicted or observed to be greater than 50 mg chl a m^{-2}, only 1 of the 10 prediction/observation points were within the RMSE confidence envelop. Another issue with this model fit analysis is also the skewness of the distribution of observed and predicted values, with most points within $1/6^{th}$ of the range of potential values (<50 mg chl a m^{-2} with a range of 0-300 mg chl a m^{-2}). Basically, it seems the model was not tested in the range of conditions in which it is intended to be applied.

3. Please comment on uncertainty in the model predictions.

The uncertainty of model predictions was problematic because: the model was not validated well for a key endpoint variable; the model was used to make predictions for nutrient conditions outside the range for which the model was calibrated and validated; and the model did not simulate extreme values well. In particular, the inability of the computer model to simulate extreme values in benthic algal biomass was a concern.

The poor prediction of algal biomass and inability to really evaluate model prediction of pH and other important response variables was discussed above.

A basic tenet of modeling, either statistical or highly calibrated computer models, is limiting extrapolation of results outside the range of conditions in which the model was developed. This model was employed outside the range of conditions for which it was calibrated. Since the computer model performed much worse when applied to September than August conditions, due to likely seasonal effects, wouldn't we also expect the same issues with performance outside the range of nutrient concentrations in which the model was calibrated?

Process based models (i.e. computer models) are theoretically better than statistical models for predicting outside the range of original conditions in which they were calibrated. However, the extent and magnitude of calibration from an initial values used in model is a key issue for using process based models to predict outside the range of calibration. Prediction outside the range of conditions for which either the statistical or process based model was calibrated requires that we know enough about the system and the behavior of the system in the two ranges of conditions (e.g. August versus September, or low and high nutrient concentrations) that we are confident that the models accurately describe behavior of the system. The less that you have to calibrate a model to new conditions to get a good fit, the more confident you can be that the model will perform well in a new set of conditions. The more fundamental the processes are that are simulated in the model and the fewer number of assumptions made for use of the model, the more certain you can be that the model will predict responses well in a set of conditions for which it was not calibrated.

Since there is little evidence that the model did perform well, either calibrating for key endpoints or predicting responses during validation, we should have concerns about accuracy of predictions by the model for ecological responses in higher nutrient concentrations for which the model was

tested. In addition, many key parameters in the model were changed greatly during calibration from what were initially thought to be appropriate. So based on model performance, we cannot be certain that it will perform well outside the range of conditions in which it was calibrated, or even within that calibration range for some key parameters.

Many assumptions needed for the model also seemed to reduce credibility of its results. Some assumptions were probably met as well in the Yellowstone River as anywhere. For example, the assumption about the model simulating a steady state equilibrium is certainly more appropriate for rivers like the Yellowstone with snow-melt dominated and relatively predictable hydroperiods versus many other rivers where storm events have dramatic and unpredictable effects on hydroperiod.

Violation of model assumptions by the ecosystem may also explain why the model simulated the ecosystem poorly. Of course assumptions are necessary, but some violations of assumptions or combinations of violations may accumulate explain the unsatisfactory behavior in the model. Here are a few examples:

- The assumption that velocity and channel substratum are "sufficiently well mixed vertically and laterally" (pg 5-8, lines 3-4) may explain why the high algal biomasses were not simulated. If average, versus optimal velocity and substratum were used, that would underestimate the high algal accrual possible in optimal velocity and substratum conditions.
- Why assume dynamic equilibrium between particle re-suspension (drift) and deposition (settling)(pg. 8-20, lines 24-25)?
- Why assume the typical meterological year during a ten year period. For example, to understand the conditions under which problems would arise 1 in 10 years, aren't regional weather patterns a likely cause of those problems. Rather than running a typical meteorological year, shouldn't the 10-year extremes be boundary conditions for a run to understand the effects of less common conditions?

In addition to violation of the assumptions in the model, there may be issues with the analytical foundation of the model to accurately represent ecosystem processes; but I am not sufficiently familiar with the model to make that judgment. For example:

- Were growth patterns and differing spatial resource limitation (density dependence) for macroalgae and microalgae or algal taxa included in the model?
- Space limitation in the model, if I understand it correctly, is not the correct conceptualization of the process that regulates density dependent growth of benthic algae. Developing a more realistic characterization of the processes regulating benthic algal accumulation and density-dependent depletion of nutrients within mats would be very interesting and perhaps improve model predictions. Effects of mixing and diffusion vary greatly between different types of algae that grow in differing nutrient and temperature ranges, such as macroalgae (*Cladophora*) and microalgae (diatoms).
- Was N-fixation included in the model and the potential for N transfer between epiphytic diatoms with cyanobacterial endosymbionts on *Cladophora*? It is possible that *Cladophora* cells close to the substratum take up nutrients and transfer them to younger, actively growing cells in the ends of the filaments suspended in the water column. Only cells at the tips of *Cladophora* filaments reproduce, so they are younger and have fewer

epiphytes than cells at the base of filaments. *Cladophora* cells that are closer to the substratum, having more epiphytes, bacteria, and entrained detritus as well as slower currents, have greater potential for uptake of recycled nutrients in the epiphytic assemblages around them than younger cells in the water column. *Cladophora* does not have complete cross walls between cells, so fluid in cells can theoretically mix between cells, which would be facilitated by the movement and bending of filaments in currents. Thus, nutrient concentrations in the water column may be poor estimators of nutrient availability to *Cladophora*, as well as other benthic algae, because of nutrient entrainment and recycling in the mats.

If many potentially important processes are not included in the model, they may either independently or cumulatively have great effects on model outcome and prediction of extreme conditions and risk of problems required for criteria development.

Another reason for questioning model predictions could be the high nitrogen and phosphorus concentrations that are predicted to generate nuisance blooms of benthic algae: 700 μg TN L^{-1} and 90 μg TP L^{-1} in Unit 3 to prevent pH violations and 1,000 μg TN L^{-1} and 140 μg TP L^{-1} in Unit 4 to prevent nuisance benthic algal problems. Although we know relatively little about nutrient concentrations affecting pH in river, these phosphorus concentrations are many times higher than phosphorus concentrations thought to cause nuisance levels of benthic algal biomass, e.g. greater than 150 mg chl a m^{-2}. Admittedly, there's a great range limiting and saturating nutrient concentrations in the literature, but a 30 μg TP/L benchmark was proposed in the Clark Fork, which is upstream from this location. Why have higher numbers in the larger mainstem of the Yellowstone River? If we assume Leibig's law of the minimum, and nitrogen and light are sufficiently great to allow algae to grow, why wouldn't the marginal habitats of the Yellowstone River generate nuisance algal biomasses at 30 μg TP/L? At least one reason could explain that discrepancy. The reactive portion of the TP may be lower in the Yellowstone River than in smaller streams where nuisance blooms of benthic algae commonly occur at TP concentrations around 30 μg TP/L. The soluble fractions of total nutrient concentrations, assumed to be the most readily available fractions, were very low in the Yellowstone River during low flow conditions (Table 6-6). However, caution should be exercised when assuming only the soluble fraction of TP is bioavailable; mounting evidence indicates that entrained particulate P and N are recycled in benthic algal mats.

The model prediction that low DO is not likely in the Yellowstone River seems reasonable. The Yellowstone River is relatively hydrologically stable, so it is probably not prone to types of extreme low flow events that allow development of low DO with resulting fish kills. Rivers and streams are probably much more susceptible to high pH and fluctuating pH conditions than to low DO; but both phenomena have not been studied sufficiently to understand thoroughly.

4. Please comment on the appropriateness of using response variables, such as chl-a and pH, as model endpoints for numeric criteria derivation, and thus protection of water quality from nutrient pollution. Please comment on the spatial application of different response variables for deriving numeric nutrient criteria (pH was used for the upstream segment while benthic algal biomass was used in the downstream segment).

pH and algal biomass response are appropriate endpoints for justification of nutrient criteria. pH is more directly linked to negative effects on aquatic fauna than nutrient concentrations, so pH is a more proximate threat to a valued ecological attribute. High algal biomass is known be an aesthetic problem in rivers, as established in the great study by Suplee et al. As described above, nutrient criteria above minimally disturbed conditions that prevent nuisance algal accumulations and violation of pH and DO standards may not protect biodiversity of some nutrient-sensitive taxa; however chl a and pH, as well as DO, are appropriate endpoints for protecting designated uses.

The most sensitive response (e.g. chl a, pH, or DO) to a stressor (i.e. nutrients in this case) should be used to establish stressor criteria, even if different response endpoints are the most sensitive in different types of habitats (in this case shallow and deep river habitats). An important goal of environmental management should be protection of ecosystem services. Of course all ecosystem services should not have to be protected in all waters, but appropriate protection is warranted. Montana DEQ and presumably a majority of the people of Montana have supported water quality criteria related to pH and benthic algae. So nutrient concentrations should not be allowed that would generate unacceptable risk of violating the pH and nuisance algal biomass criteria.

The focus on shoreline algal biomass was also appropriate because that is where people most commonly observe the water as they use the resource for recreational purposes.

5. What other analytical methods would you suggest for deriving numeric nutrient criteria for the mainstem Yellowstone River?

The appropriate methods were used to gather information about the development of nutrient criteria, but the results of the computer model were overstated and overweighted in a premature decision on nutrient criteria.

Processed based (computer) models are very informative and valuable, but they are just one line of information. Three basic research approaches can be used to develop numeric nutrient criteria: observing patterns in nature and quantifying relationships between nutrients and key endpoint variables with by statistical models (e.g. regression models); simulating patterns in nature using process-based models; and experiments in controlled environments in which environmental conditions are purposefully manipulated. Each of these methods complement each other. When they all do not agree, then conclusions are suspect. In this case, the predictions of the computer model do not match results of other research based on statistical models and experiments. Even though there are plausible reasons for those discrepancies, there is little reason that the computer model is accurate.

Despite that lack of fit between computer model predictions and measured conditions in the river, during both calibration and validation, the computer model was used. In a simple comparison of accuracy of the computer model predictions of high algal biomass as a result of higher nutrient concentrations (Figure 10-5) and the regression model characterizations between algal biomass and either TN or TP (Figure 15-2), show the regression model warranted more credibility. For the computer model, there was no relationship between algal biomass predicted and the algal biomass observed at stations (Figure 10-5). Plotting these abundances in Figure 10-5 on a log-log scale may have improved the apparent fit, but lack of fit at higher biomasses is likely. Remember the discussion above about lack of data points above 50 mg chl a m^{-2} and poor range of observed conditions. For the regression models, the results were variable but plausible (Figure 15-2). If N:P ratios are low and N limits algal growth, then we'd expect a relationship between algal biomass and TN and not between algal biomass and TP concentration. The range of TP concentrations (and bioavailable P indicated by those concentrations) may have been above the TP concentration considered to have strong effects on benthic algal growth (e.g. 30 μg TP/L). The range of TN concentrations may have crossed the sensitive range and below the limiting nutrient concentration for TN; therefore TN may have been the primary limiting nutrient in the Yellowstone River. Thus, the Montana DEQ got a relationship between TN concentrations and benthic algal biomass, but not TP concentrations and benthic algal biomass. I disagree with the interpretation by Montana DEQ about these relationships. These relationships do show that TN concentrations below 505 μg TN/L should constrain average algal biomass to less than 150 mg chl a m^{-2}, but the lack of significance in the TP algal biomass relationship indicates it should not be used to set a TP criterion. This relationship between TN and algal biomass is really the only evidence in the report for nutrient regulation of benthic algal biomass.

If benthic algal biomass is not simulated accurately by the computer model, can we trust predictions of pH and DO that respond to changes in algal biomass? pH and DO predictions of the computer model were also not validated well because of low sample sizes and ranges of conditions in which the model was calibrated.

Another question develops about whether TP concentrations need to be kept below a TP criterion that would constrain algal biomass, if TN concentrations are below that 505 μg/L; but that question is a policy deeper policy question. If TN is kept below 505 μg/L, then presumably there would not be a response of benthic algae to TP if N is the primary limiting nutrient. However, the 505 TN and 30-60 TP range seem close to what I would expect to be saturating nutrient concentrations. So, a combination of TN and TP criteria would provide double protection against risk of high algal biomass.

Good calibration of models, computer or regression, should not be expected in a river without a good range of nutrient that result in algal problems at some place across the range of nutrient conditions. In habitats in which no algal problems are observed, it is possible that sediments and low light constrain algal accumulation such that nutrients have no effect on instream algal-related conditions. In this case, downstream effects should be the concern/endpoints of criteria. Alternatively, it is possible that most that we know about the asymptotic relationship between nutrient concentrations and algal biomass is not true; or for some other reason, TP concentrations above 50-100 μg TP/L do regulate benthic algal biomass. Then the high nutrient concentrations as those proposed (700 μg TN L^{-1} and 90 μg TP L^{-1} in Unit 3 to prevent pH violations and 1,000 μg TN L^{-1}

and 140 μg TP L^{-1} in Unit 4 to prevent nuisance benthic algal problems) would be appropriate in the Yellowstone River.

Continued research in the form of monitoring of the Yellowstone River, surveys of other large rivers, experimental research, and computer modeling will be needed to develop nutrient criteria that protect ecosystem services of large rivers without overprotection. Continued monitoring in the Yellowstone River will enable assessment of whether nutrient concentrations are increasing and nuisance algal biomasses and high pH are becoming more frequent. This will forewarn managers that nutrient related problems are developing and will provide the additional information needed for better computer and regression models used to establish nutrient criteria. In the report, Montana DEQ did propose continued monitoring and data analysis with one goal being learning more about nutrient effects in the river for potential revision of the proposed nutrient criteria. But will reducing the nutrient criteria, based on new science, be practical politically. Why will the public believe the new science if the old science was not sufficient? Why hurry to have nutrient criteria if there are no known problems? Was this the wrong place to try to develop nutrient criteria for large rivers?

A concerted national effort should be developed and maintained to gather the kind of information needed for developing nutrient criteria in large rivers. Monitoring data as well as experimental results should be gathered and evaluated with statistical models and integrated in processed based models to provide sufficient information for development of nutrient criteria in large rivers. Great similarities exist among the large rivers of the world, such that information learned in multiple rivers should be able to be synthesized and related to other large rivers. Until this information is gathered and analyzed, perhaps the most prudent nutrient management strategy is to try to maintain current conditions if there are no existing problems.

A couple editorial changes worthy of note:
Figure 9-1 makes much more sense to me if Table 8-1 were changed to Table 9-1.
Figures 13-4 and 15-2 were hard to understand because the independent variable (nutrient concentration) was not on the X axis.